International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-9, Issue-5, October 2022
https://doi.org/10.55524/ijirem.2022.9.5.7
Article ID IJIR-1254, Pages 53-58
www.ijirem.org

Using Different Methodologies of Data Science to Find Comparison Between Them for Cyber-Security

Sangeeta Devi[1], Pranjal Maurya[2], Munish Saran[3], Rajan Kumar Yadav[4], and Upendra Nath Tripathi[5]

[1,2,3,4]Research Scholar Department of Computer Science, DDU Gorakhpur University, Gorakhpur, Uttar Pradesh, India
[5]Associate Professor, Department of Computer Science, DDU Gorakhpur University, Gorakhpur, Uttar Pradesh, India

Correspondence should be addressed to Sangeeta Devi; sangeeta2316@gmail.com

ABSTRACT- The main objective of this paper is that we have to find out which methodology is effective for Data Science for the cyber security problem. First of all, we discuss in the modern world, that data science is one form of topic where research spans many academic fields. It consists of scientific methods, procedures, formulas, and systems to gather information and work on that subject. When data sciences gather and store big data, analytical approaches can be used on cyber-security solutions. With the aid of a mathematical model, machine learning and big data analysis approaches can be used to manage the effects of threats. Huge amounts of data are the foundation of existing cyber-security solutions since more data allows for more accurate analysis. In data science, it is necessary to employ data analysis to resolve issues and provide answers to protect people from cybercrime projects. In this article, we compare the CRISP-DM, KDD process, and FMDS data science methodologies with their strong and weak points.

KEYWORDS- Cyber Security, CRISP-DM, KDD, FMDS.

I. INTRODUCTION

In today's quick-paced and connected society, cyber-security (CS) is one of the most critical issues. On the one hand, IoT and other computing technologies have made many discoveries for easy-going business and daily life, but on the other hand, numerous security breaches are announced every day. Both individuals and companies lost millions of dollars as a result of these security breaches. On the Internet, there are several cyber-security databases to select from. To increase cyber-security, it is necessary to take advantage of these datasets by extracting useful information. Data science (DS) and machine learning (ML) techniques can be combined to enhance cyber-security because ML techniques can be used to extract useful information from unstructured data's strengths and weaknesses.

Data science offers data-driven forecasts, which could strengthen and improve the decision-making process. Principles, procedures, and techniques are needed to comprehend a problem throughout the review and data analysis. Computer science, mathematics, and statistics are just a few of the disciplines that data science pulls from. To tackle business problems from a data perspective, data scientists must use data mining algorithms.

We will discuss different data science methodologies for cyber security protection. A huge volume of data can be used to do a more precise analysis, which could lead to the resolution of the security issue. In this paper, we discussed the outline of data science and its two key concepts, as well as an explanation of three methodologies of data science. For comparison of different approaches is presented along with a discussion of their advantages and disadvantages. We conclude this study by recommending a practical approach that might address all potential needs for cyber-security solutions.

A. Overview of Data Science

In the modern world, data science is one form of topic where research spans many academic fields. It consists of scientific methods, procedures, formulas, and systems to gather information and work on that subject. This area spans a variety of genres and serves as a common hub for the integration of machine learning, data analysis, and statistics. Real-time data, machine learning techniques, and theoretical knowledge of statistics all come together in this to produce high-yielding results for the company. The Economic Times reports that demand for data science personnel in India has increased by more than 400 percent across a range of business sectors. In today's world, we use different techniques employed in data science, and can implicit better decision making, which otherwise humans might be miss eye and mind. Remember, but the machine never forgets it. Artificial intelligence (specifically machine learning) has become an important tool for data scientists.[1]

Data science is a set of basic ideas that enables the rational extraction of knowledge and information from data. Identifying patterns, values, and user interests, is quite similar to data mining, which attempts to extract this information using technologies and apply and utilize it for relationship management and behavior analysis. The industries around the globe are moving due to data science. Because "security is all about data," it is very crucial for the development of brilliant cyber-security systems and services.

B. Overview of cyber security

Cyber security technology is a process by which computers, networks, servers, mobile devices,

electronic systems and data are adopted to protect against hostile intrusion. Its other name is also information technology security or electronic information security. Cyber security systems include protecting networks, programs, devices, and data by using technology and processes and controls to protect against cyber attacks. Therefore, its main purpose is to connect systems, networks, programs and data [2]. Table 1 shows the threats types.

C. The Importance and Benefits of Cyber Security

- It can protects the business against data violation and cyber-attacks.
- It can prevent access of unauthorized users.
- It has improved the time of recovery after a violation.
- Regulation of observance
- This can create mutually positive trust in the reputation of the company among all stakeholders (customers, partners, developers and workers).
- It protects end-user and endpoint device security.
- It also protects the data and network.
- It also helps to continue the business.
- Different Types of threats for cyber security

D. Types of Threats to Cyber Security

There are different types of threats in Cyber Security:

Table 1: Different types of threats

Malware	Malware is harmful software it harms any file or is programmed to be used against a computer. Worms, viruses, Trojan horses, and spyware are included.
Phishing	Phishing is social attack software. They send fraudulent emails or text messages to steal their sensitive data or information such as credit card or login information, etc.[2].
Distributed-Denial-of-Service (DoS)	A denial of service (DoS) preventing from responding to requests like attack overloads on a computer or network and it preventing from responding requests.
Man-in-the-middle-attack	When hackers interject themselves into a two-party transaction, a man-in-the-middle (MITM) attack takes place.
Ransomware	Malware can also include ransomware. It entails an attacker encrypting and locking the targeted computer system files, then demanding to decrypt the code and unlock the system.
Cloud vulnerabilities	Hackers have more options for data decryption and unlocking data stored and transferred online.
manage to poor data	It's easier to lose and expose important data and information when we have a huge amount of data.
Poor cyber hygiene	It is very essential to train an employees who access the network and maintain safe cyber practices,.

E. Concept of Data Science with Cyber Security

In Cyber security, there are two types of decisions for data science:

- Data discovery decisions based on and
- Frequent decision-making processes decisions.

Data discovery is the process of gathering complex data and turning it into manageable information for users. Data scientists can use their expertise to identify potential threats and protect against attacks in the realm of cyber security. Data discovery are the process of gathering complex data and turning it into manageable information for users. Data analysis and information appraisal are essential to routine decision-making.[4]

To obtained the final result, raw data are processed in various phases. In the first phase, collected data are prepared for analysis. In the next phase of creating models utilizing various data analysis methods.

II. METHODOLOGY

A methodology is a broad strategy that directs the methods and actions within a particular domain. The process doesn't rely on certain tools or technologies. A methodology, on the other hand, provides a framework for obtaining outcomes while utilizing a variety of techniques, procedures, and heuristics. [5]

Beneficial knowledge should be extracted from data using systematic processes and procedures that follow predetermined phases. The knowledge that may be extracted from data must be carefully considered and the results must be assessed from the perspective of being utilized. This is because the knowledge can be used to help with decision-making in a specific application. It is generally advantageous to break the business problem down into parts that correspond to computing or estimating or evaluate the values and estimating probabilities, along with a structure for rejoining the parts. [6]

One of the data science topics that should be taken into account in connection to cyber security is correlation discovery. Usually, it gives information on data items that shed light on other data items, especially recognized quantities that lessen the uncertainty of unknown numbers. CRISP-DM, SAS SEMMA, FMDS, and KDD process are only a few of the approaches available for solving cyber-Security and data science problems,[8]

A. CRISP-DP (Cross-Industry Standard Process for Data Mining

In 2014, CRISP-DM remained the most popular approach for data mining projects, accounting for 42% of all data mining initiatives. Other methodologies for data mining and data science challenges include SAS SEMMA, KDD process, and CRISP-DM. 7.5% more people are using the KDD Process.[5]

Although Rollins emphasizes several new data science techniques, such as the use of big data, the integration of text analytics into predictive modeling, and process automation, he also displays a core methodology that is comparable to CRISP-DM, SEMMA, and KDD Process. Additionally, Microsoft presents the Team Data Science Process (TDSP), which proposes a life cycle for data science initiatives.

This study compares the basic framework for data science with the KDD Process, CRISP-DM, TDSP, and other approaches (FMDS). Although SAS SEMMA and TDSP are not included in this analysis due to a significant fall in applications, FMDS, KDD, and CRISP-DM have been chosen since they are thought to be the most popular. Because it offers beginning and fundamental needs for knowledge discovery, the KDD Process has also been included. The reason TDSP was not chosen is that it is tailored for projects involving artificial intelligence or machine learning that are closely related to cyber security applications [8]. Figure 1 shows stages of CRISP-DM.

The six phases of CRISP-DP are as follows:

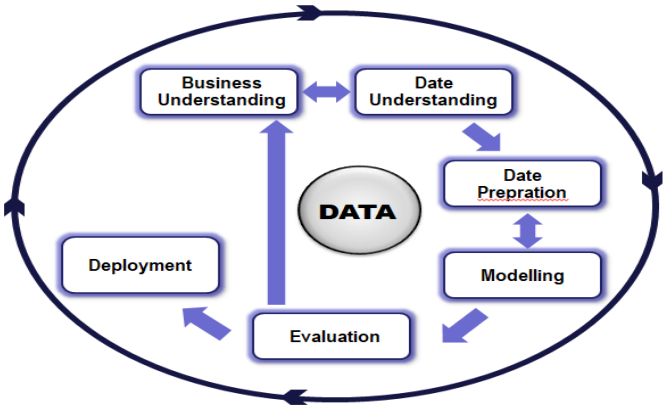

Figure 1: Displays the Six Stages. Of CRISP-DM

Numerous case studies have employed CRISP-DM, and it is fairly well documented. The success of the project depends on CRISP-systematic, DM's well-defined, and well-documented approach being independent of data mining tools. The CRISP-DM makes perfect sense and appears to be common sense. Because using a widely used methodology improves quality and efficiency, many techniques and cutting-edge analytic platforms are built on the CRISP-DM stages. For even the most complex data science projects of the present, according to Vorhies, CRISP-DM offers solid direction [6]. Table 2 represents phases for CRISP-DM

Table 2: Phases of CRISP-DM

SNO.	Phases	Description
1	Business Understanding	It's intended to concentrate on comprehending the project or problem's objectives and requirements from a business standpoint.
2.	Data Understanding	This phase starts with the initial data gathering, followed by tasks to familiarize oneself with the data and identify issues with data quality.
3.	Data Preparation	The entirety of the work necessary to create the final dataset from the first raw data is covered in this step.
4.	Modeling	In this phase modeling techniques and strategies are selected and implemented, identifying their specific parameters and pre-requisites.
5.	Evaluation	During this phase, modeling approaches and tactics are chosen, and used their particular requirements and conditions should be noted.
6	Deployment	This comprises all data preparation and processing processes necessary to handle raw data to provide a final output that is consistent with the model's development process.

B. KDD (Knowledge Discovery in Database)

Utilizing any necessary preprocessing, sampling, or data transformation procedures, is the way of using data mining techniques to extract knowledge based on specific metrics and thresholds in a database. The application domain perception must also be taken into account when developing, enhancing, and improving the KDD process. [7] Figure 2 shows KDD life cycle and Table 3, Table 4 shows phases for KDD and phases for FMDS respectively.

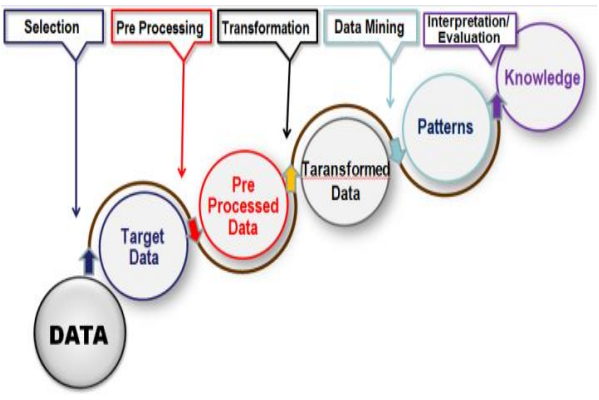

Figure 2: Phases of KDD Life Cycle

Table 3: Different Phases of KDD are as Follows

S. No.	Phases	Descriptions
1	Selection	It refers to creating a target data set, focusing on a data samples or particular subset of variables in a database.
2	Pre-Processing	This phase aims to obtain consistent data.
3	Transformation	In this stage, feature dimensionality is reduced by utilizing data transformation techniques.
4	Data Mining	This step should involve attempting to identify attention or behavioral trends utilizing a particular set of data mining tools.
5	Evaluation	The last stage should involve evaluating and interpreting the mined pattern.

C. FMDS (Foundation Methodology for Data Science)

To deliver massive data sets, text analysis, and image analysis, data scientists use FMDS process. It includes the following actions [8]. Figure 3 represents the phases of FMDS.

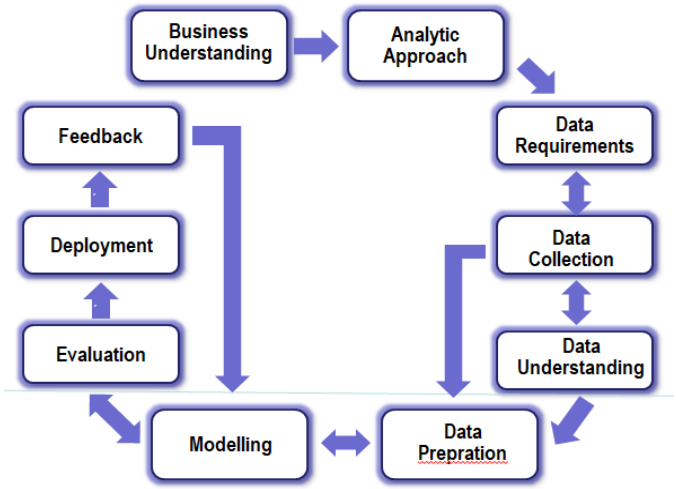

Figure 3. Different Phases of FMDS

Table 4: Different Phases of FMDS are as follows

S. No	Phases	Description
1	Business Understanding	It's intended to concentrate on comprehending the project or problem's objectives and requirements from a business standpoint.
2	Analyses approach	In this stage, a machine learning technology is used to pinpoint an appropriate analytical strategy after the issue has been recognized.
3	Data Collection	It is an important phase to locate and collect accessible data sources that are pertinent to the issue domain at this initial data gathering phase.
4	Data Requirement	The necessary data needs are established in this phase when an acceptable approach has been determined.
5	Data understanding	This phase starts with the initial data gathering, followed by tasks to familiarize oneself with the data and identify issues with data quality.
6	Data preparation	The entirety of the work necessary to create the final dataset from the first raw data is covered in this step.
7	Modeling	In this phase modeling techniques and strategies are selected and implemented, identifying their specific parameters and pre-requisites.
8	Evaluation	During this phase, modeling approaches and tactics are chosen and used, and their particular requirements and conditions should be noted.
9	Deployment	This comprises all data preparation and processing processes necessary to handle raw data in order to provide a final output that is consistent with the model's development process.
10	Feedback	This phase is the last phase for gathering results from the analytical model's implemented version in order to evaluate and provide feedback on its functionality, performance, and usefulness in light of the deployment environment.

III. ANALYSIS AND COMPARISON

When comparing CRISP-DM and KDD, we found that CRISP-DM includes the business understanding and implementation stages, but KDD does not.
Users must apply real data in a commercial setting, so deployment is also crucial. This stage has the potential to turn knowledge into a cyber security issue.
Three data science methodologies are compared in this report. KDD is only a technique for obtaining knowledge; it is not entirely ideal. Because they understand the necessity for a repeatable strategy, the majority of analytic managers employ the CRISP-DM process, although there are certain issues with this technique as it is typically used. Four issues of CRISP-DM like as Poor rework, blind handoffs to IT, weak iteration, and poor iteration [7] [8]. Table 5 represents methodologies for data science.

Table 5: Comparison Table of Three Data Science Methodologies

KDD	CRISP-DM	FMDS
	Business Understanding	Business Understanding
		Analytic Approach
		Data Requirements
Selection	Data understanding	Data Collection
Pre Processing		Data understanding
Transformation	Data Preparation	Data Preparation
Data mining	Modeling	Modeling
Interpretation and Evaluation	Evaluation	Evaluation
	Deployment	Deployment
		Feedback

In this study, three data science approaches are contrasted. KDD is not a completely ideal methodology, it is merely a method for knowledge and learning. The majority of analytical managers adopt the CRISP-DM method since they are aware of the need for a repeatable strategy, although there are certain problems with this approach as it is frequently applied. The four problems with CRISP-DM are poor rework, blind handoffs to IT, weak iteration, and poor iteration.
The advantages of FMDS should be taken into account for contemporary cyber-security projects are as [9]:

- FMDS is a practical tool that works across platforms. The entire analytical procedure can be run to make sure the analyzed outcomes are appropriate in all circumstances, not just a sample modeler. By testing these models, cyber-security projects can become more effective and dependable.
- The model process is accelerated and refreshed automatically to produce better outcomes.

In both KDD and CRISP-DM, the evaluation, deployment, and feedback phases have the potential to be superior to the straightforward evaluation phase. The FMDS's feedback phase has the potential to generate creative queries that can improve cyber-security projects and add new features. As FMDS is more broad and independent of any platform tool, it cuts down on the time needed for data preparation for the authenticity and dependability of cyber-security problems.

Three data science methodologies are compared in this report. KDD is only a technique for obtaining knowledge; it is not entirely ideal. Because they understand the necessity for a repeatable strategy, the majority of analytic managers employ the CRISP-DM process, although there are certain issues with this technique as it is typically used. Poor rework, blind handoffs to IT, weak iteration, and poor iteration are the four issues of CRISP-DM. [10][11].

IV. CONCLUSION

The data science approach describes the general steps. In this essay, we evaluate three methodologies of data science for the issue of cyber-security. All of the steps are included in FMDS. In light of the study provided, we conclude that FMDS includes all favorable aspects of a cyber-security project. Additionally, it is tool and platform independent. Because it is properly planned out with clearly defined stages, it might be suggested in any cyber-security project. In this paper, we describe the three well-known methodologies of data science: CRISP-DM, KDD, and FMDS, and also explain the relationship between them. After comparison, we find that FMDS is the best methodology among them.[10][12].

REFERENCES

[1] Sarker IH et al (2020) Cyber security data science: an overview from machine learning perspective. J Big Data 7(1):1–29
[2] https://www.comparitech.com/vpn/cybersecurity-cyber-crime-statistics-facts-trends/
[3] Galeano-Brajones J et al (2020) Detection and mitigation of DoS and DDoS attacks in IoT-based stateful SDN: an experimental approach. Sensors 20(3):816
[4] Meidan Y et al (2018) N-baiot—network-based detection of iot botnet attacks using deep autoencoders. IEEE Pervasive Comput 17(3):12–22
[5] https://analyticsindiamag.com/top-10-datasets-for-cybersecurity-projects/
[6] Shahzadi S et al. (2021) Machine learning empowered security management and quality of service provision in SDN-NFV environment
[7] Yener B, Gal T (2019) Cyber security in the era of data science: examining new adversarial models. IEEE Secur Priv 17(6):46–53
[8] Farhad Foroughi and Peter Luksch, "Data Science Methodology for Cyber security project" https://www.researchgate.net> publication
[9] Humayun M et al (2020) Cyber security threats and vulnerabilities: a systematic mappingstudy. Arab J Sci Eng 1–19
[10] Wikipedia, "Cross Industry Standard Process for DataMining," http://en.wikipedia.org/wiki/Cross Industry_Standard_Process, http://the-modeling-agency.com/crisp-dm.pdf
[11] "Introduction To Hadoop –NYOUG", https://nyoug.org>SIG>DataWarehousing
[12] The DETER Project - Advancing the Science of Cyber Security Experimentation and Test. Terry Benzel, Jelena Mirkovic, et al. IEEE HST 2010 Conf, Boston, MA, November 2010.

ABOUT THE AUTHORS

Sangeeta Devi received the Master of Computer Application (MCA) from IGNOU New Delhi and Master of Technology (M.Tech.) from AKTU Lucknow. She is currently Ph.D. research Scholar in the Department of Computer Science, DDU Gorakhpur University. Her research interest includes Data Science, WSN, IoT, Machine Learning and Deep Learning.

Pranjal Maurya received the Bachelor of Technology (B.Tech.) in Computer Science Engineering of Technology & Management and Master of Technology (M.Tech.) in Computer Science Engineering (CSE) from Madan Mohan Malaviya University of Technology. She is currently Ph.D. research Scholar in the Department of Computer Science, DDU Gorakhpur University. Her research interest includes WSN, Cloud Computing, IoT, Machine Learning and Deep Learning. She was previously working in Institute of Technology & Management as Assistant Professor for 1 years.

Munish Saran received the Bachelor of Technology (B.Tech.) in Computer Science Engineering (CSE) from Babu Banarasi Das National Institute of Technology & Management and Master of Technology (M.Tech. Gold Medal) in Computer Science Engineering (CSE) from Madan Mohan Malaviya University of Technology. He is currently Ph.D. research scholar in the Department of Computer Science, DDU Gorakhpur University. His research interest includes Cloud Computing, IoT, Machine Learning and Deep Learning. He was previously working in Infosys as senior system engineer for 4 years.

Rajan Kumar Yadav received the Bachelor of Science (B.Sc.) in computer Science from Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur (Uttar Pradesh, India) and Master of Computer Application (MCA) from Madan Mohan Malaviya University of Technology. He is currently Ph.D. Research Scholar in the Department of Computer Science, DDU Gorakhpur University. His Research interest includes Cloud Computing, Machine Learning and IoT.

Dr. Upendra Nath Tripathi is currently Associate Professor in the Department of Computer Science, Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur. He has 21 years of teaching and research experience. His areas of interests are Database, IoT, Machine Learning, Cloud Computing and Data Science.